



MATNLI MA'LUMOTLARNI KENGAYTIRISH UCHUN DINAMIK STRATEGIYA REJALASHTIRISHNI O'RGANISH MASALASI

Subxonqulov Umidjon

Buxoro davlat universiteti

subxonqulovumidjon@gmail.com

Dolzarblik

Chuqur o'rganish modellarining yuqori samaradorligi asosan katta hajmdagi belgilangan ma'lumotlarga bog'liq bo'lib, ma'lumotlarni qo'lda belgilash jarayoni yuqori xarajat va vaqt talab etadi. Bu holat cheklangan hajmdagi ma'lumotlarga ega vazifalarda modelning qayta moslashuvi (overfitting) xavfini kuchaytiradi. Tabiiy tilni qayta ishlash (NLP) sohasida ma'lumotlarni kengaytirish (data augmentation) ushbu muammoni yumshatishning samarali vositasi sifatida e'tirof etilgan. Biroq mavjud kengaytirish usullarining aksariyati statik, ya'ni butun o'qitish jarayoni davomida o'zgarmas strategiyadan foydalanadi. Amalda esa turli vazifalar va ma'lumotlar to'plamlari uchun optimal kengaytirish operatsiyalari farqlanadi, shuningdek, bitta modelning o'qitish jarayonining boshlang'ich va yakuniy bosqichlarida ham turlicha strategiya talab etilishi mumkin. Shu sababli, o'qitish davomida dinamik ravishda moslashuvchan kengaytirish siyosatini (policy) avtomatik shakllantirish masalasi zamonaviy NLP tadqiqotlarida dolzarb ilmiy muammo hisoblanadi.

Tadqiqot maqsadi

Tadqiqotning asosiy maqsadi - matnli ma'lumotlarni kengaytirish jarayonida o'qitish davri (epoch)ga bog'liq holda o'zgaruvchan kengaytirish siyosatini avtomatik izlab topuvchi LPST (Learning Policy Scheduling for Text augmentation) usulini ishlab chiqishdan iborat. Bunda quyidagi vazifalar qo'yiladi: (1) so'z darajasidagi bir necha parametrlashtirilgan kengaytirish





operatsiyalarini o'z ichiga olgan izlanish maydonini (search space) shakllantirish; (2) ushbu maydonda o'qitish davri bo'yicha eng samarali gipoparametrlar ketma-ketligini (policy schedule) topish uchun populyatsiyaga asoslangan o'qitish algoritmidan foydalanish; (3) taklif etilgan yondashuvni matn tasniflash va mashina tarjimai vazifalarida sinab ko'rish orqali uning umumlashtirilgan samaradorligini isbotlash.

Metodologiya

Tadqiqotda qo'llanilgan yondashuv quyidagi asosiy komponentlardan iborat.

1. Gipoparametr izlash sifatida siyosat shakllantirish. Kengaytirish funksiyasi $f(S; \theta)$ statik ko'rinishdan $f(S; \theta, t)$ dinamik ko'rinishga o'tkaziladi, bunda t joriy o'qitish davrini bildiradi. Kengaytirish funksiyasi differensiallanmaydigan xarakterga ega bo'lgani sababli, gradientga asoslangan optimallashtirish o'rniga populyatsiyaga asoslangan o'qitish (Population Based Training, PBT) algoritmi qo'llaniladi. Bir nechta yordamchi model parallel ravishda o'qitiladi va har ikki davrdan so'ng ekspluatatsiya-tadqiqot (exploit-explore) jarayoni amalga oshiriladi: eng past natijali modellar eng yaxshi natijali modellarning gipoparametrlari va og'irliklarini nusxalab oladi (ekspluatatsiya), so'ngra gipoparametrlar tasodifiy siljiriladi yoki qayta tanlanadi (tadqiqot).

2. Kengaytirish operatsiyalari izlanish maydoni. Uchta asosiy so'z darajasidagi operatsiya taklif etiladi: (a) Almashtirish (Replacement) — so'zni uning atrofidagi so'zlar, maxsus belgi (placeholder) yoki lug'atdan tasodifiy tanlangan so'z bilan almashtirish; (b) Almashinuv (Swap) — jumladagi ikki so'zning o'rnini almashtirish; (v) Kengaytirish (Growth) — jumla boshiga atrofdagi yoki tasodifiy so'zlarni qo'shish orqali jumlani uzaytirish. Har bir operatsiya qo'llash ehtimoli (p), qo'llanish miqyosi (m), nishon so'zga masofa (s) va ichki quyi-siyosatlar taqsimoti (w) kabi to'rtta gipoparametr bilan tavsiflanadi.



3. Baholash protokoli. O'qitish ma'lumotlarining 20 foizi tekshiruv (validation) to'plami sifatida ajratiladi. O'n oltita yordamchi model 200 davr davomida o'qitiladi, eng yaxshi natijaga erishgan modelning har bir davrdagi gipoparametrlari qayd etilib, yagona dinamik siyosat jadvali (policy schedule) hosil qilinadi. So'ngra ushbu jadval asosida yangi model qaytadan o'qitiladi va test to'plamida baholanadi. Yondashuv samaradorligi CNN, ikki yo'nalishli LSTM va Transformer arxitekturalarida, beshta matn tasniflash va ikkita mashina tarjimai (WMT'14, WMT'17 ingliz-nemis) vazifalarida, mavjud statik kengaytirish usullari (Drop, Swap, Blank, Replace, SwitchOut, EDA) bilan qiyoslanadi.

Asosiy natijalar

Tadqiqot natijalari LPST usulining barcha uchta neyron arxitektura (CNN, Bi-LSTM, Transformer) bo'yicha ko'pchilik ma'lumotlar to'plamlarida statik kengaytirish usullaridan ustunligini ko'rsatdi. Xususan, CNN arxitekturasida beshta ma'lumotlar to'plamining to'rttasida eng yuqori aniqlik ko'rsatkichiga erishildi, o'rtacha aniqlik esa asl (kengaytirilmagan) ma'lumotlarga nisbatan sezilarli darajada oshdi. Mashina tarjimai vazifalarida ham LPST WMT'14 va WMT'17 to'plamlarida BLEU ko'rsatkichi bo'yicha barcha bazaviy usullardan yuqori natija ko'rsatdi. Muhim topilma shundaki, statik siyosatlarni (faqat boshlang'ich yoki faqat yakuniy davrda topilgan gipoparametrlarni) butun o'qitish jarayoniga tatbiq etish hatto kengaytirilmagan ma'lumotlarga nisbatan ham pastroq natija berdi — bu dinamik siyosat rejalashtirishning zarurligini bevosita tasdiqlaydi. Gipoparametrlar vizualizatsiyasi shuni ko'rsatdiki, turli operatsiyalarning optimal ehtimol qiymatlari o'qitish davri davomida sezilarli darajada o'zgaradi va bu o'zgarish ma'lumotlar to'plamiga qarab farqlanadi. Ablatsion tahlil uchta operatsiyaning (Almashtirish, Almashinuv, Kengaytirish) har biri yakuniy samaradorlikka ijobiy hissa qo'shishini, hech biri ortiqcha emasligini tasdiqladi.



Ilmiy yangilik: matnli ma'lumotlarni kengaytirish siyosatini birinchi marta o'qitish davriga bog'liq dinamik jarayon sifatida shakllantirish va uni gipoparametr izlash masalasiga keltirib, populyatsiyaga asoslangan o'qitish algoritmi orqali samarali yechish; bunda kengaytirish siyosatini izlash xarajati yagona modelni o'qitish xarajatiga deyarli teng bo'lib qoladi.

Amaliy ahamiyati: taklif etilgan yondashuv cheklangan hajmdagi (10 mingdan kam namunali) korpuslarga ega matn tasniflash va mashina tarjimasi tizimlarida, xususan resurs jihatdan tanqis tillar uchun modellarni o'qitishda qo'shimcha belgilangan ma'lumot talab qilmasdan model sifatini oshirish imkonini beradi, bu esa amaliy NLP tizimlarini ishlab chiqishda xarajatlarni kamaytirishga xizmat qiladi.

Xulosa

O'tkazilgan tadqiqot natijasida matnli ma'lumotlarni kengaytirish masalasida statik siyosatlardan dinamik, o'qitish davriga moslashuvchan siyosatga o'tishning ilmiy va amaliy asosliligi isbotlandi. Kengaytirish siyosatini gipoparametr izlash muammosi sifatida qayta talqin qilish va uni populyatsiyaga asoslangan samarali algoritm yordamida yechish orqali qo'shimcha hisoblash xarajatlarisiz sezilarli sifat yaxshilanishiga erishildi. Kelgusi tadqiqotlar izlanish maydonini kengaytirish, jumla darajasidagi murakkabroq operatsiyalarni joriy etish, shuningdek taklif etilgan dinamik rejalashtirish tamoyilini boshqa NLP vazifalariga, jumladan matnni generatsiya qilish va so'zlashuv tizimlariga tatbiq etish yo'nalishida davom ettirilishi maqsadga muvofiqdir.

FOYDALANILGAN ADABIYOTLAR

1. Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., et al. (2017). Population based training of neural networks. arXiv preprint arXiv:1711.09846.



2. Wei, J. W., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. Proceedings of EMNLP-IJCNLP.
3. Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., & Le, Q. V. (2018). AutoAugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501.
4. Ho, D., Liang, E., Chen, X., Stoica, I., & Abbeel, P. (2019). Population based augmentation: Efficient learning of augmentation policy schedules. International Conference on Machine Learning, 2731–2741.
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 5998–6008.
6. Wang, X., Pham, H., Dai, Z., & Neubig, G. (2018). SwitchOut: An efficient data augmentation algorithm for neural machine translation. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 856–861.
7. Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. arXiv preprint arXiv:1805.06201.
8. Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1715–1725.
9. Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1532–1543.
10. Li, S., Ao, X., Pan, F., & He, Q. (2022). Learning policy scheduling for text augmentation. Neural Networks, 145, 121–127.